Mapping rat probes from Agilent, Illumina and Affymetrix expression arrays, in view of comparing titration data in the MAQC2 experiment.

Danielle and Jean Thierry-Mieg, October 2, 2008

To compare the titration results obtained on the three platforms, we need to identify the sets of microarray probes that measure the same genes, or even the same transcript(s) within the genes. The first step is to annotate the rat genes, the second to map the probes, and the third to generate the compatible sets of probes across the platforms.

1. Annotation of Rattus norvegicus genes

The NM/NR RefSeq represent a good start, but their aim is to provide one good model per conserved protein coding genes: they do not try to give a comprehensive representation of the transcriptome and its flurry of alternative variants, or non-conserved and non-protein-coding genes. So they mainly ignore ESTs or Traces, and sometimes even GenBank.

The XM/XR predictions, added to the RefSeq set, and similarly the Ensembl set have a large component of predicted models, usually based on conservation and not driven by cDNA evidence; some are pure ab initio predictions. In large transcriptome studies, the fit of predicted models without cDNA evidence to the experimental results is low.

On the other hand, AceView aims at a comprehensive representation of the known transcriptome, irrespective of conservation or protein coding potential. We align all cDNA sequences available on the genome and quality control cDNA clones for common artefacts (strand inversion, partial deletion of the insert or internal priming). We then reconstruct transcript models by grouping the experimental cDNAs sequences into a minimal number of contiguous compatible groups. Finally, we group transcript models sharing intron boundaries or having substantial sequence overlap into genes. All AceView transcripts have cDNA support and 92% of all experimental cDNA sequences are represented in AceView. As expected, the fit to the large scale transcriptome studies is improved, although still incomplete.

On September 13, 2008, we downloaded the ~1 million rat cDNA sequences from NCBI:

- 15,358 RefSeq NM/NR (as usual, we excluded the XM/XR)
- 4,341 mRNAs from GenBank,
- 837,339 ESTs from dbEST,
- 99,241 sequences from the NCBI Trace repository [downloaded September 26]

We reconstructed 27,202 main genes, of which 23,128 are spliced and 4,704 are single exon genes potentially encoding good proteins. Many of the main AceView genes are annotated in Entrez Gene (19,966 main genes have at least one geneID, they actually account for 20,219 GeneIDs because 268 genes contain two or more RefSeq models defining separate genes in Entrez). However, remarkably, 7,236 novel main genes supported by cDNAs sequences (26% of the total) are not yet in Entrez Gene. Of those, 7,057 are spliced genes, a proof that they are transcribed.

In addition to the main genes, we reconstruct 14,109 unspliced possibly partial genes (we call ‘putative’) and 60,817 ‘cloud’ genes (fragmons if you wish).

The 23,128 spliced genes include 45,026 (alternatively) spliced mRNAs, so we currently annotate only 2 alternative variants per spliced rat gene on average, versus 5 for mouse and 8 for human. Actually, the current rat transcriptome sequencing is still sparse; in AceView genes, we integrate about 8 million human sequences and 5 million mouse sequences, versus less than 0.8 million for rat.

Nevertheless the genes look good, we still have to collect more disease associations from RGD, but we have made this database public October 4th! Please check www.aceview.org and click on the nice rat picture, then type in any rat gene name, or words, or mRNA accession or even nothing and ‘Go’.

2. Mapping probes to the reference genome and to various transcriptomes (RefSeq NM/NR, all RefSeq including the XM/XR, Ensembl and AceView)

We are sorry to confirm that the rat genome is still only of draft quality: when we allow for a maximum of 2 mismatches per probe, and map microarray probes for rat human and mouse to their current respective NCBI reference genome, the percent of mapped probes averages 84% in rat, versus 90% in human and 93.3% in mouse. When we map to the recent RefSeq NM/NR, the percentage of mapped probes averages 50.0% in rat, versus 56.3% for human and 65.7% in mouse. We therefore downloaded the new 2008 genome from Baylor which exploits the recent whole genome shotgun sequencing traces, but found it did not improve probe mapping (as expected, there is a bias and this genome may still be far better).

So we allowed for up to 3, 6 and 7 base mismatches for Affy, Illumina and Agilent respectively (where one mismatch is a single base deletion, insertion, transition or transversion) in the probe-to-reference-genome or probe-to-transcriptome alignment.

For the rat transcriptome, we used 4 sets of RNA models:

- the recent RefSeq NM/NR (downloaded Sept 13, 2008)
- the entire set of RefSeq including predicted models (NM/NR/XM/XR from Sept 13, 2008)
- the entire set of Ensembl models (downloaded September 23, 2008)
- the AceView transcripts and genes (summarizing public cDNAs as of September 13, 2008)

For genome we used the NCBI RefSeq reference genomes for the three species, and also the Baylor 2008 new rat genome. Human and mouse are listed at the same stringency for comparison.

Here are the results of probe mapping, allowing for 3|6|7 mismatches, when we ignore the strand (mapping on sense or antisense of transcripts is considered equivalent).

Species	Mapping on sense or antisense of mRNA models Array	Probes on the array	Aligning to RefSeq NM/NR	Aligning to RefSeq NM/NR/ XM/XR	Aligning to Ensembl	Aligning to AceView	Aligning to the reference genome	Aligning to Baylor genome 2008
Rat	Affy.Rat230_2	342410	148846	184138	139102	309850	309673	306655
	Agilent.Rat	41011	21210	27924	25051	33670	33889	33197
	Illumina_RatRef-12_V1_0_R3	22523	13288	17532	17337	17324	19715	19051
Mouse	Affy_Mouse430_2	496468	270698	301884	293355	466800	466904	
	Illu.MouseWG-6_V1_1_R3	46632	26553	30307	30755	42935	43983	
	Illu.MouseWG-6_V2_0_R1	45281	31056	34359	34302	42383	42904	
Human	Affy_U133_Plus	604258	323183	354129	354970	573526	576872	
	Affy_encode	732045	44518	51696	67180	137731	732045	
	AgilentWHG	41000	26937	29975	30017	38345	36689	
	Illumina.human_MAQC1	47282	20827	23527	27834	37206	44276	
	Illumina.HumanWG-6_V2_0_R3	48701	23087	26544	25188	42368	43825	
	Illumina.HumanHT-12_V3_0_R1	48803	28434	31852	30353	43977	44815	

And now in percentage

	Mapping on sense or antisense of mRNA models Array	Probes	% to NM/NR RefSeq	% to any RefSeq, NMNR /XM/XR	% to Ensembl	% to AceView transcripts	% to reference genome	% to 2008 rat genome (BCU)
Rat	Affy.Rat230_2	342410	43.5%	53.8%	40.6%	90.5%	90.4%	89.0%
	Agilent.Rat	41011	51.7%	68.1%	61.1%	82.1%	82.6%	80.0%
	Illumina_RatRef-12_V1_0_R3	22523	59.0%	77.8%	77.0%	76.9%	87.5%	84.0%
Mouse	Affy_Mouse430_2	496468	54.5%	60.8%	59.1%	94.0%	94.0%	
	Illu.MouseWG-6_V1_1_R3	46632	56.9%	65.0%	66.0%	92.1%	94.3%	
	Illu.MouseWG-6_V2_0_R1	45281	68.6%	75.9%	75.8%	93.6%	94.8%	
Human	Affy_U133_Plus 2	604258	53.5%	58.6%	58.7%	94.9%	95.5%	
	Affy_encode	732045	6.1%	7.1%	9.2%	18.8%	100.0%	
	AgilentWHG	41000	65.7%	73.1%	73.2%	93.5%	89.5%	
	Illumina.human_MAQC1	47282	44.0%	49.8%	58.9%	78.7%	93.6%	
	Illumina.HumanWG-6_V2_0_R3	48701	47.4%	54.5%	51.7%	87.0%	90.0%	
	Illumina.HumanHT-12_V3_0_R1	48803	58.3%	65.3%	62.2%	90.1%	91.8%	

We notice that Illumina in human is getting closer to AceView, except that there appears to be a strand problem in the latest design (see tables below). The AceView mRNA models are available for any use on our ftp site ☺.

Now mapping on sense only, at the same stringency

Species	Mapping on sense of mRNA models Array	Probes	Aligning to RefSeq NM/NR	Aligning to RefSeq NM/NR /XM/XR	Aligning to Ensembl	Aligning to AceView	Aligning to the reference genome	Aligning to Baylor genome 2008
Rat	Affy.Rat230_2	342410	141954	173808	130001	298936	309673	306655
	Agilent.Rat	41011	17258	23061	21592	26548	33889	33197
	Illumina_RatRef-12_V1_0_R3	22523	13230	17441	17212	17094	19715	19051
Mouse	Affy_Mouse430_2	496468	264596	293538	285174	448074	466904	
	Illu.MouseWG-6_V1_1_R3	46632	26247	29854	30330	42443	43983	
	Illu.MouseWG-6_V2_0_R1	45281	30764	33963	33901	42014	42904	
Human	Affy_U133_Plus	604258	302818	330303	331182	541539	576872	
	Affy_encode	732045	21167	25444	33083	72088	732045	
	AgilentWHG	41000	25375	28253	28330	35652	36689	
	Illumina.human_MAQC1	48701	23000	26264	24752	34520	43825	
	Illumina.HumanWG-6_V2_0_R3	48803	28347	31612	29963	37852	44815	
	Illumina.HumanHT-12_V3_0_R1	47282	20677	23173	27287	33871	44276	

	Mapping on sense of mRNA models PERCENT Array	Probes	% to NM/NR RefSeq	% to any RefSeq, NMNR/XM	% to Ensembl	% to AceView transcripts	% to reference genome	% to 2008 rat genome (BCU)
Rat	Affy.Rat230_2	342410	41.5%	50.8%	38.0%	87.3%	90.4%	89.0%
	Agilent Rat	41011	42.1%	56.2%	52.6%	64.7%	82.6%	80.0%
	Illumina_RatRef-12_V1_0_R3	22523	58.7%	77.4%	76.4%	75.9%	87.5%	84.0%
Mouse	Affy_Mouse430_2	496468	53.3%	59.1%	57.4%	90.3%	94.0%	
	Illumina.MouseWG-6_V1_1_R3	46632	56.3%	64.0%	65.0%	91.0%	94.3%	
	Illumina.MouseWG-6_V2_0_R1	45281	67.9%	75.0%	74.9%	92.8%	94.8%	
Human	Affy_U133_Plus 2	604258	50.1%	54.7%	54.8%	89.6%	95.5%	
	Affy_encode	732045	2.9%	3.5%	4.5%	9.8%	100.0%	
	AgilentWHG	41000	61.9%	68.9%	69.1%	87.0%	89.5%	
	Illumina.human_MAQC1	48701	47.2%	53.9%	50.8%	70.9%	90.0%	
	Illumina.HumanWG-6_V2_0_R3	48803	58.1%	64.8%	61.4%	77.6%	91.8%	
	Illumina.HumanHT-12_V3_0_R1	47282	43.7%	49.0%	57.7%	71.6%	93.6%	

3. Selection of groups of probes measuring the same genes, or sets of transcripts:

We generated groups of probes when the three platforms hit the exact same gene or sets of genes, or the exact same transcript or sets of transcripts. To generate a group, we demand at least 8 probes for Affy, one for Agilent and one for Illumina. By construction, the probes in the groups hit a unique set of gene(s) or transcript(s) at the chosen stringency, so this guarantees they are not ambiguous. A nice feature however is that if two or more repeated genes (or transcripts) are so related in sequence that some probes from all platforms cannot distinguish them, the probes would - as is desirable - still be listed as a group, but would hit a set of multiple genes (or transcripts).

We identified groups of corresponding probes across platforms relative to the following transcriptome models (downloaded in September 08):

- The 15,348 NM/NR RefSeqs RNA models.
- All 35,343 RefSeq (i.e. NM and NR, and also the predicted models XM/XR).
- The entire Ensembl transcriptome, including 82,261 models.
- All AceView models from main genes, i.e. 56,994 models, of which 45,026 are spliced. These are computed In two flavors, either through the best tested model or by demanding that the probes touch exactly the same set of transcripts.
- Finally, we calculated the groups in a gene-centric way, generating the list of genes measured uniquely by probes from all three platforms.

It might be interesting to compare the results on the six (or 12) sets!

Groups at Gene level :

Depending on the granularity of the experiment, it may be advantageous to simply use groups of probes measuring the same gene(s). That is a good choice if quantitative variations are expected to dominate over the finer grain differential alternative splicing or polyadenylation. At the level of the gene, there are 10,407 groups measuring a total of 10,594 genes (10,288 groups measuring 10,682 genes if we lost the strand information) tested by all three platforms with good probes. 166 groups hit more than one gene, altogether measuring 353 closely repeated genes (371 if strand is lost, altogether measuring 765 genes).

Groups measuring the best tested AceView transcript per gene

Here we extended the notion that drives RefSeq and applied it to AceView transcripts. AceView includes RefSeq models as if they were cDNAs, but in that file we automatically gain genes for which no RefSeq exists or the RefSeq is not tested by the 3 platforms, but another transcript is (e.g. , Aebp2, where RefSeq supports variant a, but all probes are designed against variant b, which is the only variant supported by cDNAs...). This set is a subset of the gene set, which ensures that probes are non-ambiguous. Furthermore, most often the test is done on variant a (the most protein coding in our nomenclature). But when variant b or c is used, only probes touching b without touching a are reported, so there is some specific filtering added here.

Additionally, we take into account the ‘validated alternative polyadenylation sites’ that we have been annotating over the last year. We won’t describe the algorithm in detail, but we provide in the files the position of the 3’ ends, along the transcript indicated, the number of polyadenylated cDNAs ending there (i.e. the support) and the sequence of the polyA signal seen at the expected position.

You may use these groups directly, or if you care for alternative polyadenylation, just avoid going across a 3’ end line (3p3p3p), especially if supported by many cDNAs. Affymetrix frequently tests alternative 3’ ends. For information, in human there are on average 3.5 such alternative sites per gene, and there is evidence that choice of polyadenylation sites is strongly regulated and important biologically, as it is involved in both microRNA regulation and transcript stability control.

Groups measuring the exact same sets of AceView transcripts:

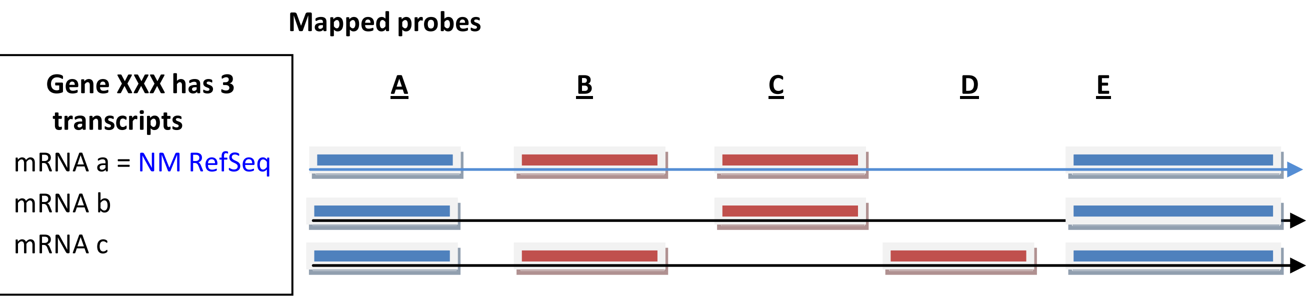
If one is interested in finer grain analysis, where both alternative splicing and alternative polyadenylation matters, this is the list of choice. All probes listed here contact exactly the same set of variants and no other. This set should give the most coherent measures across platforms. However it may be too stringent because if a variant is partial (its sequence is not completely known at one end), lack of knowledge will create a somewhat artificial edge and ruin an otherwise good group, that will not make it into this list.

As in the best tested transcript file, we indicate the polyadenylation sites, here in the longest RNA.

Groups measuring RefSeq or Ensembl

The definition of groups of probes depends on the annotation of the genes. The richer the transcriptome annotations, the more stringent and difficult to meet the group definition becomes. There is hopefully a balancing effect: when a more complete transcriptome is considered, more probes map to transcripts. For instance in AceView we map more probes to genes than in RefSeq NM/NR (+18% to 77% in Illumina, +30% to 82% in Agilent and +47% to 91% in Affymetrix in the sense-independent mapping which applies in our case).

The diagram below shows a gene with 3 alternative mRNA variants, the top variant (a) is the RefSeq; blue exons are constitutive, red exons are alternative. Five probes (or probesets) A, B, C, D, E measure this gene’s expression, and map as depicted. All 5 probes measure the same gene, only probes A and E measure the same set of variants (a, b and c), but probes A, B, C and E measure the same RefSeq transcript.



The numbers of probe groups in each case is given in the table below and the lists are in the attached zipped files. Numbers are given first under the hypothesis that strand information has been lost (as seems to be the case), then if the strand was not lost. For completeness because this may be useful for other studies as well, we have prepared the same files for human and mouse (unfortunately no Agilent there as we do not have the probe sequences).

Platforms tested to touch the very same set of transcripts	strand	NM/NR	All RefSeq NM/NR XM/XR	All Ensembl	**Same set of AceView transcripts**	**Best tested AceView transcript**	**AceView genes tested by all platforms**
RAT Affy 230_2 at least 8 probes Agilent at least 1 probe Illu Ref-12 V1.0.R3 at least 1 pr	sense	8,538	9,348	3,365[1]	**7,038**	**9,401**	**10,407**
	any	8,673	9,552	3,808	**7,273**	**9,306**	**10,288**
MOUSE (No Agilent, only Affy et Illumina) Affy 430_2 at least 8 probes Illumina WG6 V2 0 R1 at least 1	sense	15,358	16,076	11,983	**14,464**	**17,919**	**17,480**
	any	15,340	16,073	11,942	**13,852**	**17,119**	**17,115**
HUMAN Affy U133 Plus 2 at least 8 probes Agilent WHG at least 1 Illumina HT 12 V3 0R1 at least 1	sense	15,992	15,809	10,089	**9,277**	**14,869**	**16,295**
	any	15,972[3]	15,757[3]	10,112	**8,980[3]**	**14,042**	**15,739**

[1] Notice that despite the very large number of Ensembl rat models, probes do not match them well.

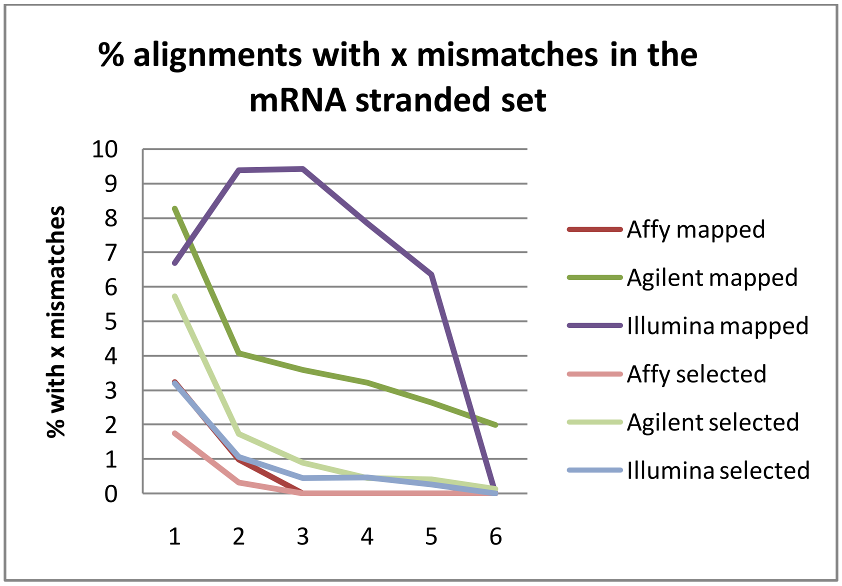
[3] comparing the strand-dependent versus sense or antisense strand results, one should not be surprised to see some numbers decrease: you gain the cases where a company has designed the probe on the 'wrong' strand, but you may lose in some cases where there are genes in antisense in a given region and one company has a probe in the overlap (and may see variants from both genes) and the other does not (and sees variants from only one gene).

Note on mapping stringency

We allowed for many mismatches because of the draft nature of the rat genome, but we provide the number of mismatches in the table, in case you prefer to lower the thresholds. Here are the (interesting) statistics on number of mismatches,

		Number of alignments with N mismatches							
	Platform	0	1	2	3	4	5	6	7
after mapping	Affymetrix	298925	293064	9694	2984	0	0	0	0
	Agilent	26548	23655	2195	1082	952	856	702	530
	Illumina	17094	15790	1142	1603	1610	1340	1084	0
after selecting groups (stranded, all mRNAs)	Affymetrix	93055	91143	1620	292	0	0	0	0
	Agilent	8946	8110	512	155	80	40	37	12
	Illumina	7232	6839	231	77	32	34	19	0

And a little diagram, in percent, showing that our selection criteria automatically reduced the % mismatches!



Format of the attached files

The twelve rat files (NM/NR, all RefSeq, Ensembl, AceView set of transcripts, AceView best transcript, AceView gene, each either strand specific or not) have the following format:
Col 1: Group number
Col 2: type of line: either ‘Target’ (transcript(s) or gene(s)) or ‘Probe’ or ‘3p end’ . Then information in the next columns depend on the content here . This allows us to document everything useful for the selected group of probes in a single blob, easily readable even by eye, in the table.

‘Target’ is the transcript or sets of transcripts, or the gene or genes measured in common specifically by the group of probes. When more than one target is measured (for instance multiple mRNAs) they come as a space delimited list in column 5.
‘Probe’ is the identifier of the probe sequence on the microarrays (tell us if you prefer another prefix)
‘3p end’ validated polyadenylation site.

Col 1	Probe group number		
Col 2 is the word:	Target (transcript or groups of transcripts or genes)	Probe	3p end
Col 3 : coordinate or length, in nucleotide All coordinates are in order along the gene: neighbors in the list are neighbors in the transcript or gene, 3’ ends appear at their actual position too	length of target mRNA or gene indicated in column 4	coordinate of probe in target mRNA or gene indicated in column 4, in increasing order	coordinate of validated 3’ end in target mRNA or gene in column 4
Col 4: name of transcript or gene relative to which the coordinate in column 3 is given (when multiple, we pick the longest one)	name of mRNA or gene target	Name of mRNA or gene target	Name of mRNA or gene target
Col 5:	Complete name of target or targets (there may be multiple transcripts or genes, then just space delimited)	name of probe (A for Agilent Rat for Affymetrix ILMN for Illumina)	3p3p3p3p
Col 6	number of Affymetrix probes in the group. Often there are more than 8	Number of single base mismatches relative to reference genome	Number of cDNA supporting clones ending at this 3’ end site; indicates support strength
Col 7:	number of Agilent probes in the group		Type of polyA signal (standard is AATAAA, any single letter variant can be used)
Col 8:	number of Illumina probes in the group		

Note the probes come in order along the gene, so your best chances of reducing impact of alternative 3’ ends is to pick the most compact block of probes including Agilent Illumina and the number you like from Affymetrix probes (we could provide a reduced group view, with for instance 8 Affy probes. Because these are selected and filtered through, and all ambiguous probes are removed, 8 probes is certainly plenty to get a reliable signal.

If you want the human and mouse files, just ask and we will send them as well.
Enjoy!
Danielle and Jean